Representation of missense variants for predicting modes of action

Guojie Zhong
Department of Systems Biology
Columbia University
New York, NY 10032
gz2294@cumc.columbia.edu

Yufeng Shen
Department of Systems Biology and Biomedical Informatics
Columbia University
New York, NY 10032
ys2411@cumc.columbia.edu

Abstract

Accurate prediction of functional impact for missense variants is fundamental for genetic analysis and clinical applications. Current methods focused on generating an overall pathogenicity prediction score while overlooking the fact that variant effect should be multi-dimensional via different modes of action, such as gain or loss of function, and loss of folding stability or enzymatic activity. Recent breakthrough of high-capacity language models enabled *ab initio* prediction of protein structures as well as self-supervised representation learning of protein sequence and functions. Here we present RESCVE, a method to learn universal representation of sequence variation from protein context. We demonstrated the utility of the method predicting a range of modes of action for missense variants through transfer learning.

1 Introduction

Missense variants are the most common and major type of coding variants that contribute to many human diseases[1, 2, 3, 4, 5]. In the past decade, many methods have been developed to predict functional effects of missense variants. Traditional methods, such as GERP, Polyphen2, SIFT, CADD, REVEL, M-CAP, Eigen and MPC[6, 7, 8, 9, 10, 11, 12, 13], utilized manually encoded features including sequence conservation, local protein structure properties, population allele frequency with conventional machine learning models like SVM and random forests to output a single score that reflect pathogenicity. Methods like PrimateAI and gMVP showed deep learning frameworks like convolutional neural networks (CNN) and graph attention could capture protein context and co-evolutional strength from sequence and multiple sequence alignment, which improved performance[14, 15]. Moreover, recent protein language models based on Transformers and self-supervised training on billions of protein sequences in UniProt[16] showed great success in protein structure prediction[17], context representations[18] and zero-shot predictors of variant pathogenicity[19]. However, those methods were not optimized for distinguishing modes of action of pathogenicity. Pathogenic missense variants could act through different modes to cause disease and may result in markedly different clinical phenotypes[20]. Generally, pathogenic missense variants could disturb protein function in two ways, Gain of Function (GoF) and Loss of Function (LoF). For example, GoF variants could result in hyper-activity, altered selectivity in ion channel genes[21, 22] as well as constitutive activation, increased sensitivity and lower specificity in signaling proteins[23]. Likewise, LoF variants could disable protein function in different modes, such as decreased net ion flow in ion channel genes[21], decreased protein stability[24], decreased enzymatic activity[25] or loss of interaction domain[26, 27, 28]. More specifically, the mode of action is about how the normal function of a protein is perturbed by the mutation. Since different proteins have different functions, it would be conceptually ambiguous to build a generic model that directly predict missense variants modes of action for all proteins, with the exception of fold stability. In this study, we aim to build a model that

Machine Learning for Structural Biology Workshop, NeurIPS 2022.

learns universal representation of coding sequence variants and use transfer learning to predict modes of action in each protein or protein family (equation 1).

$$\begin{aligned} h &= f_0(AA_{ref}, AA_{alt}, X) \\ S_i &= f_i(h), i \in Protein\ Families \end{aligned} \tag{1}$$

Here we present RESCVE (REpresentation of protein Sequence Context for Variant functional Effect prediction), which utilized the protein language models and pre-trained on ClinVar/HGMD datasets[29, 30] to obtain a latent representation for variants. We showed such representation could characterize modes of action in multiple tasks, including gain and loss of function in ion channel proteins, protein stability and enzyme activities of PTEN with state-of-the-art performances.

2 Build latent representations of variant effect based on protein language models

2.1 Overview of the model

Most of the recent protein language models were trained with the mask-and-predict task, which is self-supervised to maximize the predicted probability of reference amino acid at a given position given the protein sequence context. Such pre-training settings allowed zero-shot prediction of variant effect by comparing the probability of mutated amino acid with reference amino acid, which showed good performances in spearman correlation with deep mutational scan scores and AUC in clinical data[31, 19]. However, such metric is impossible, by definition, to distinguish different modes of action for pathogenic variants.

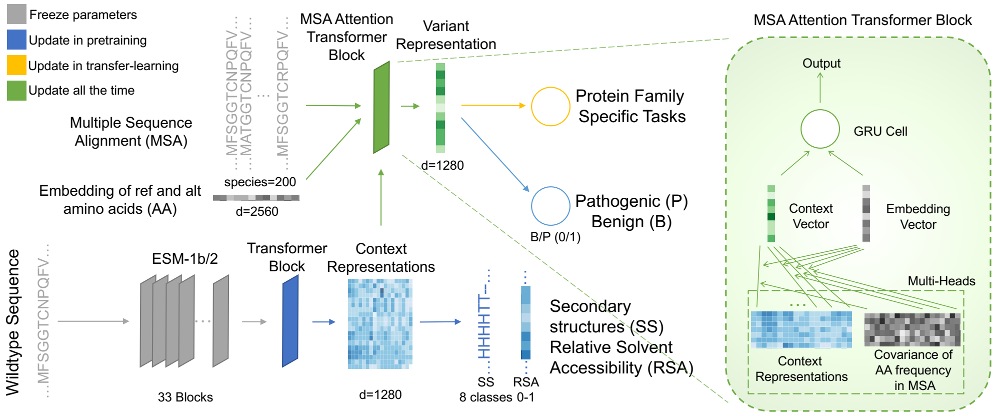

Figure 1: Model Architectures

Here we would like to take the advantage of protein language models' powerful latent representations of protein properties while build a generalized representation model for variant effects, which can be further used for transfer-learning in protein family specific tasks. The embeddings of protein language models consist of information of amino acid properties, protein structure and indirect conservation information. Secondary structures and relative solvent accessibility (RSA) are causative factors of missense variants functional impact[27, 32]. To take advantage of such prior knowledge while minimizing losing generality from the language models, we added a transformer block after the last layer of language model. The transformer block has same architecture as RoBERTa and ESM layers[33, 34, 19] and serves both a predictor for secondary structures, relevant solvent accessibility (RSA) and the input for the MSA-attention layer. The MSA-attention layer is designed to capture the impact of single amino acid change given the protein context information and multiple sequence alignment (MSA) information inspired from gMVP[15]. The protein context information is weighted summed by the attention score to the variant position embeddings, which contains both embeddings for reference and substituted amino acids. The attention score contains both query-key dot product as well as the co-evolution covariance score calculated from the MSA of 200 species (equation 2).

$$\begin{aligned}
X &= ESM(wild\,type\,sequence)\\
C &= Transformer(X)\\
Q &= W_q \cdot GELU(Linear(AA_{ref}, AA_{alt}))\\
K &= W_k \cdot GELU(Linear(C))\\
V &= W_v \cdot GELU(Linear(C))\\
A &= Q \cdot K^T + W_A \cdot tanh(Linear(cov(MSA)))\\
Attn &= Linear(softmax(A) \cdot V)\\
h &= GRU(Attn, C)
\end{aligned} \tag{2}$$

$cov(MSA)$ is a covariance matrix of shape $21 \times 21 \times seq\_len$ and was adapted from gMVP[15] and defined as below (equation 3), where A and B denotes amino acid identities (20+gap), M denotes species, i denotes variant position, j denotes indexes of other positions.

$$cov(MSA)_{ij}^{AB} = \frac{1}{M}(\sum_{m=1}^{M} \delta_{A,X_{i,m}} \delta_{B,X_{j,m}} - \sum_{m=1}^{M} \delta_{A,X_{i,m}} - \sum_{m=1}^{M} \delta_{B,X_{j,m}}) \tag{3}$$

The attention result and the embeddings of variant position were passed to a GRU cell, end up with an output of variant representation with 1280 dimensions (equation 2). A simple 3-layer multi-layer perceptron (MLP) was added to the representation with activation layer depending on the modes of action tasks (Appendix A.1 Table 3).

2.2 Data sets and tasks

2.2.1 Pre-training: Pathogenicity, Secondary Structure and Relevant Solvent Accessibility

To make the model able to recognize protein properties as well as identifying variant effects, we pre-trained the model with two sub-tasks. The first one is to predict protein secondary structures and relevant solvent accessibility (RSA), using predicted structure of 21,052 human proteins from AlphaFold2[17] and use DSSP[35, 36] annotated secondary structures as training set and 2,341 proteins as testing set. The second task is to classify pathogenic variants, using labeled data including 58,888 pathogenic or likely pathogenic variants from ClinVar[29], HGMD[30] and 56,850 benign variants from PrimateAI[14]. To reduce overfitting and avoid memorization of protein identity instead of amino acid change, we randomly added 3% (2038) of benign variants that locate in the same position as pathogenic variants (Appendix A.2 Table 4). For testing dataset, we used pathogenic variants in cancer hotspots from a recent study[37] and benign variants from PrimateAI. We crop each sequence to a length of 1001 to fit the maximum input length of ESM models. The multi-sequence alignment (MSA) files were downloaded from ensembl website[38]. The performance is quantified by the AUC of secondary structure prediction, Pearson corelation coefficient (R) of relavant solvent accessibility (RSA) and the AUC of pathogenicity classification.

2.2.2 Gain / Loss of function for 4 protein families

We curated datasets that consist modes of action of 4 protein families from Heyne, *et al*[21] and Bayrak, *et al*[20]. The gain and loss of function variants were labeled based on the clinical phenotypes and literature text mining. We randomly split it into 4:1 of training and testing datasets (Appendix A.2 Table 4). The goal of the task is to distinguish gain of function and loss of function variants. The performance is quantified by AUC of classification in the testing dataset.

2.2.3 Deep mutational scan datasets for 5 proteins

We obtained 5 deep mutational assay datasets of gene *PTEN*, *NUDT15*, *CCR5*, *CXCR4*, *VKORC1* from MAVEDB[39], each contains multiple measurements on different functional effects of single missense variants, including protein stability[24], enzyme activity[25], antibody binding, etc. We randomly split it into 8:1:1 of training, validation and testing sets (Appendix A.2 Table 4). The goal is to build a regression model from the variant representations. The performance is quantified by the Pearson correlation coefficient (R) in the testing dataset.

3 Results

3.1 Secondary Structure prediction and variant pathogenicity prediction

We trained RESCVE (Appendix A.3, A.4) and tested the performance of secondary structure and pathogenicity prediction of our model against ESM1b[18] and ESM2 (650M) model[40]. For secondary structure and RSA tasks, we trained ESM models with a simple MLP layer on same training dataset. For pathogenicity task, we use the likelihood ratio of amino acid change as zero-shot prediction for ESM[19]. For our model RESCVE, we compared four settings: ESM2 (650M)-based, ESM2 (15B)-based, ESM1b-based and ESM1b-based structure without transformer layer (Appendix A.5). ESM1b-based RESCVE outperforms the others in the pathogenicity task while reached decent performances in secondary structure and RSA tasks (Table 1). Thus we decided to use ESM1b-based RESCVE for transfer learning.

Table 1: pre-train task performances

Name	Secondary Structure (AUC)	RSA (R)	Pathogenicity (AUC)
ESM1b	0.751	0.784	0.922 (zero-shot)
ESM2 650M	0.657	0.556	0.815 (zero-shot)
ESM1b based RESCVE	0.952	0.919	**0.951**
ESM1b based RESCVE (no Transformer)	0.798	0.844	**0.951**
ESM2 650M based RESCVE	0.951	0.910	0.940
ESM2 15B based RESCVE	**0.970**	**0.923**	0.908

3.2 Modes of action tasks

We tested the performance of several protein family specific transfer learning tasks with our model against three base-line models: zero-shot prediction of ESM1b, a simple elastic net model that trained on same training dataset using ESM1b's embeddings, ESM1b-based RESCVE without the transformer layer. The RESCVE models performed better in most tasks, indicating that our model structure can efficiently extract useful information in language model's embeddings for modes of action prediction (Table 2). For protein families like Ion channel and SH2, our model has AUC above 0.9, while for other protein families, our model has AUC around 0.70 (Table 2), which could be related to the limited sample sizes (Appendix A.2 Table 4). We did notice that removing the transformer layer could result in slightly better performances in some of the tasks. Generally, removing the transformer layer will decrease model's ability of transfer learning and the model will do better at tasks that preferred by original language model (i.e. in *PTEN*, *NUDT15*, *VKORC1*). We further explored the transfer learning ability of RESCVE with limited data points. The result showed that RESCVE could adapt to most tasks with only 60$\%$ of dataset (Appendix A.6). Finally, we showed that our model can potentially benefit clinical diagnosis and treatments by distinguishing different functional effects of single variants (Appendix A.7).

4 Related Work

Multiple models have been trained to predict variant pathogenicity from different aspects. Including supervised annoted-feature based models GERP, Polyphen2, SIFT, CADD, REVEL, M-CAP, Eigen, MPC[6, 7, 8, 9, 10, 11, 12, 13]; supervised deep learning models PrimateAI, gMVP[14, 15]; unsupervised deep learning models ESM1v[19], EVE[41], Tranceptron[42]. There are very few methods that focused on modes of action prediction. gMVP has revealed that supervised pre-training on variant pathogenicity could help distinguish GoF/LoF variants in ion channel genes. Bayrak *et al* built a database of GoF/LoF variants based on literature search on ClinVar / HGMD[20]and developed a method to predict GoF/LoF variants across all genes using manually annotated features and gradient boosting tree algorithm[43].

Table 2: Modes of action tasks performances

Task Name		ESM1b	ESM1b	RESCVE	RESCVE
		(zero-shot)	(elastic-net)	(no Transformer)	
GoF/LoF (AUC)	Ion Channel	0.541	0.822	0.873 ± 0.008	**0.920** ± 0.006
	Kinase	0.565	0.696	0.730 ± 0.022	**0.788** ± 0.030
	SH2	0.600	0.827	0.813 ± 0.013	**0.929** ± 0.020
	cNMP binding	0.653	**0.840**	0.520 ± 0.023	0.627 ± 0.053
PTEN (R)	stability	0.458	0.622	0.636 ± 0.005	**0.666** ± 0.000
	enzyme activity	0.548	0.551	**0.724** ± 0.002	0.715 ± 0.000
NUDT15 (R)	stability	0.545	0.613	0.784 ± 0.000	**0.807** ± 0.003
	enzyme activity	0.596	0.608	**0.767** ± 0.008	0.757 ± 0.002
CCR5 (R)	stability	0.446	0.392	**0.606** ± 0.000	0.593 ± 0.002
	bind AB-2D7	0.397	0.434	**0.573** ± 0.000	0.568 ± 0.002
	bind HIV-1	0.420	0.416	0.591 ± 0.000	**0.612** ± 0.001
CXCR4 (R)	stability	0.347	0.393	**0.595** ± 0.000	**0.595** ± 0.001
	bind CXCL12	0.183	0.225	0.451 ± 0.000	**0.453** ± 0.001
	bind AB-12G5	0.220	0.264	0.367 ± 0.000	**0.393** ± 0.001
VKORC1 (R)	stability	0.622	0.242	**0.797** ± 0.000	0.724 ± 0.012
	enzyme activity	0.311	0.132	0.221 ± 0.015	**0.313** ± 0.005

5 Discussion

Application of new methods from the natural language processing field to protein biology has enabled modeling protein context using the entire collection of protein sequences. In this project, we aimed at predicting modes of action of missense variants using protein language models. We first defined several modes of action prediction tasks based on public available datasets. We note that modes of action are often specific to protein families. Gain of function mutations in kinases usually have different molecular mechanisms compared the ones in ion channels. While decrease of folding stability is a loss of function mechanism universal to all proteins, most proteins have additional specific ways for loss of function, such as reduction of the enzyme activity or perturbation of interaction interface with other proteins. Therefore, we argue that prediction of modes of action should largely be protein family specific tasks. With RESCVE, we trained a unified variant effect representation model that utilized the embeddings from latest protein language models. We pre-trained this representation model on a large set of pathogenic and benign variants and compared three language models (ESM1b, ESM2 650M, ESM2 15B) in the pre-training test, in which ESM1b-based RESCVE performs better in pathogenicity prediction while ESM2 15B-based RESCVE better in structure-related predictions. Our observation reveals the potential trade-offs between static structure prediction and variant effect prediction for protein language models, similar as others' findings[44]. Finally, we showed that this pre-trained model can be applied to multiple protein family specific modes of action prediction tasks through transfer learning. A limitation of the study is that we only tested 9 modes of action tasks. The generalizability of the methods to other protein families and modes of action has to be tested in future studies with larger and more comprehensive functional readout data. This work could be applied to more comprehensive genetic analysis as well as personalized clinical applications.

6 Future Work

We expect further development of models that represent protein sequence and structure will enable improvement in predicting modes of action for pathogenic variants. The growing data from deep mutational scan assays, in both breadth (more genes) and depth (more aspects of protein functions), will facilitate performance assessment and transfer learning training process. Finally, we will apply the method in clinical settings, including diagnostic analysis and potentially newborn screening.

Acknowledgments and Disclosure of Funding

We acknowledge Dr. Mohammad AlQuarashi, Dr. Wendy Chung and all members of Shen lab and Chung lab for useful suggestions. This work was supported by NIH R01GM120609 and Columbia University Precision Medicine Joint Pilot Grants Program.

References

[1] J. Homsy, S. Zaidi, Y. Shen, J. S. Ware, K. E. Samocha, K. J. Karczewski, S. R. DePalma, D. McKean, H. Wakimoto, J. Gorham, S. C. Jin, J. Deanfield, A. Giardini, J. Porter, G. A., R. Kim, K. Bilguvar, F. Lopez-Giraldez, I. Tikhonova, S. Mane, A. Romano-Adesman, H. Qi, B. Vardarajan, L. Ma, M. Daly, A. E. Roberts, M. W. Russell, S. Mital, J. W. Newburger, J. W. Gaynor, R. E. Breitbart, I. Iossifov, M. Ronemus, S. J. Sanders, J. R. Kaltman, J. G. Seidman, M. Brueckner, B. D. Gelb, E. Goldmuntz, R. P. Lifton, C. E. Seidman, and W. K. Chung, "De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies," *Science*, vol. 350, no. 6265, pp. 1262–6, 2015.

[2] K. L. Huang, R. J. Mashl, Y. Wu, D. I. Ritter, J. Wang, C. Oh, M. Paczkowska, S. Reynolds, M. A. Wyczalkowski, N. Oak, A. D. Scott, M. Krassowski, A. D. Cherniack, K. E. Houlahan, R. Jayasinghe, L. B. Wang, D. C. Zhou, D. Liu, S. Cao, Y. W. Kim, A. Koire, J. F. McMichael, V. Hucthagowder, T. B. Kim, A. Hahn, C. Wang, M. D. McLellan, F. Al-Mulla, K. J. Johnson, N. Cancer Genome Atlas Research, O. Lichtarge, P. C. Boutros, B. Raphael, A. J. Lazar, W. Zhang, M. C. Wendl, R. Govindan, S. Jain, D. Wheeler, S. Kulkarni, J. F. Dipersio, J. Reimand, F. Meric-Bernstam, K. Chen, I. Shmulevich, S. E. Plon, F. Chen, and L. Ding, "Pathogenic germline variants in 10,389 adult cancers," *Cell*, vol. 173, no. 2, pp. 355–370 e14, 2018.

[3] S. C. Jin, J. Homsy, S. Zaidi, Q. Lu, S. Morton, S. R. DePalma, X. Zeng, H. Qi, W. Chang, M. C. Sierant, W. C. Hung, S. Haider, J. Zhang, J. Knight, R. D. Bjornson, C. Castaldi, I. R. Tikhonoa, K. Bilguvar, S. M. Mane, S. J. Sanders, S. Mital, M. W. Russell, J. W. Gaynor, J. Deanfield, A. Giardini, J. Porter, G. A., D. Srivastava, C. W. Lo, Y. Shen, W. S. Watkins, M. Yandell, H. J. Yost, M. Tristani-Firouzi, J. W. Newburger, A. E. Roberts, R. Kim, H. Zhao, J. R. Kaltman, E. Goldmuntz, W. K. Chung, J. G. Seidman, B. D. Gelb, C. E. Seidman, R. P. Lifton, and M. Brueckner, "Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands," *Nat Genet*, vol. 49, no. 11, pp. 1593–1601, 2017.

[4] F. K. Satterstrom, J. A. Kosmicki, J. Wang, M. S. Breen, S. De Rubeis, J. Y. An, M. Peng, R. Collins, J. Grove, L. Klei, C. Stevens, J. Reichert, M. S. Mulhern, M. Artomov, S. Gerges, B. Sheppard, X. Xu, A. Bhaduri, U. Norman, H. Brand, G. Schwartz, R. Nguyen, E. E. Guerrero, C. Dias, C. Autism Sequencing, P.-B. C. i, C. Betancur, E. H. Cook, L. Gallagher, M. Gill, J. S. Sutcliffe, A. Thurm, M. E. Zwick, A. D. Borglum, M. W. State, A. E. Cicek, M. E. Talkowski, D. J. Cutler, B. Devlin, S. J. Sanders, K. Roeder, M. J. Daly, and J. D. Buxbaum, "Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism," *Cell*, vol. 180, no. 3, pp. 568–584 e23, 2020.

[5] X. Zhou, P. Feliciano, C. Shu, T. Wang, I. Astrovskaya, J. B. Hall, J. U. Obiajulu, J. R. Wright, S. C. Murali, S. X. Xu, L. Brueggeman, T. R. Thomas, O. Marchenko, C. Fleisch, S. D. Barns, L. G. Snyder, B. Han, T. S. Chang, T. N. Turner, W. T. Harvey, A. Nishida, B. J. O'Roak, D. H. Geschwind, S. Consortium, J. J. Michaelson, N. Volfovsky, E. E. Eichler, Y. Shen, and W. K. Chung, "Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes," *Nat Genet*, vol. 54, no. 9, pp. 1305–1319, 2022.

[6] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nat Methods*, vol. 7, no. 4, pp. 248–9, 2010.

[7] H. Carter, C. Douville, P. D. Stenson, D. N. Cooper, and R. Karchin, "Identifying mendelian disease genes with the variant effect scoring tool," *BMC Genomics*, vol. 14 Suppl 3, p. S3, 2013.

[8] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou, "Identifying a high fraction of the human genome to be under selective constraint using gerp++," *PLoS Comput Biol*, vol. 6, no. 12, p. e1001025, 2010.

[9] N. M. Ioannidis, J. H. Rothstein, V. Pejaver, S. Middha, S. K. McDonnell, S. Baheti, A. Musolf, Q. Li, E. Holzinger, D. Karyadi, L. A. Cannon-Albright, C. C. Teerlink, J. L. Stanford, W. B. Isaacs, J. Xu, K. A. Cooney, E. M. Lange, J. Schleutker, J. D. Carpten, I. J. Powell, O. Cussenot, G. Cancel-Tassin, G. G. Giles, R. J. MacInnis, C. Maier, C. L. Hsieh, F. Wiklund, W. J. Catalona, W. D. Foulkes, D. Mandal, R. A. Eeles, Z. Kote-Jarai, C. D. Bustamante, D. J. Schaid, T. Hastie, E. A. Ostrander, J. E. Bailey-Wilson, P. Radivojac, S. N. Thibodeau, A. S. Whittemore, and W. Sieh, "Revel: An ensemble method for predicting the pathogenicity of rare missense variants," *Am J Hum Genet*, vol. 99, no. 4, pp. 877–885, 2016.

[10] I. Ionita-Laza, K. McCallum, B. Xu, and J. D. Buxbaum, "A spectral approach integrating functional genomic annotations for coding and noncoding variants," *Nat Genet*, vol. 48, no. 2, pp. 214–20, 2016.

[11] K. A. Jagadeesh, A. M. Wenger, M. J. Berger, H. Guturu, P. D. Stenson, D. N. Cooper, J. A. Bernstein, and G. Bejerano, "M-cap eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity," *Nat Genet*, vol. 48, no. 12, pp. 1581–1586, 2016.

[12] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants," *Nat Genet*, vol. 46, no. 3, pp. 310–5, 2014.

[13] K. E. Samocha, J. A. Kosmicki, K. J. Karczewski, A. H. O'Donnell-Luria, E. Pierce-Hoffman, D. G. MacArthur, B. M. Neale, and M. J. Daly, "Regional missense constraint improves variant deleteriousness prediction," *bioRxiv*, p. 148353, 2017.

[14] L. Sundaram, H. Gao, S. R. Padigepati, J. F. McRae, Y. Li, J. A. Kosmicki, N. Fritzilas, J. Hakenberg, A. Dutta, J. Shon, J. Xu, S. Batzoglou, X. Li, and K. K. Farh, "Predicting the clinical impact of human mutation with deep neural networks," *Nat Genet*, vol. 50, no. 8, pp. 1161–1170, 2018.

[15] H. Zhang, M. S. Xu, X. Fan, W. K. Chung, and Y. Shen, "Predicting functional effect of missense variants using graph attention neural networks," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 1017–1028, 2022.

[16] C. UniProt, "Uniprot: the universal protein knowledgebase in 2021," *Nucleic Acids Res*, vol. 49, no. D1, pp. D480–D489, 2021.

[17] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[18] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proc Natl Acad Sci U S A*, vol. 118, no. 15, 2021.

[19] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives, "Language models enable zero-shot prediction of the effects of mutations on protein function," *bioRxiv*, p. 2021.07.09.450648, 2021.

[20] C. Sevim Bayrak, D. Stein, A. Jain, K. Chaudhary, G. N. Nadkarni, T. T. Van Vleck, A. Puel, S. Boisson-Dupuis, S. Okada, P. D. Stenson, D. N. Cooper, A. Schlessinger, and Y. Itan, "Identification of discriminative gene-level and protein-level features associated with pathogenic gain-of-function and loss-of-function variants," *Am J Hum Genet*, vol. 108, no. 12, pp. 2301–2318, 2021.

[21] H. O. Heyne, D. Baez-Nieto, S. Iqbal, D. S. Palmer, A. Brunklaus, P. May, C. Epi, K. M. Johannesen, S. Lauxmann, J. R. Lemke, R. S. Moller, E. Perez-Palma, U. I. Scholl, S. Syrbe, H. Lerche, D. Lal, A. J. Campbell, H. R. Wang, J. Pan, and M. J. Daly, "Predicting functional effects of missense variants in voltage-gated sodium and calcium channels," *Sci Transl Med*, vol. 12, no. 556, 2020.

[22] H. A. Lester and A. Karschin, "Gain of function mutants: ion channels and g protein-coupled receptors," *Annu Rev Neurosci*, vol. 23, pp. 89–125, 2000.

[23] Y. X. Tao, "Constitutive activation of g protein-coupled receptors and diseases: insights into mechanisms of activation and therapeutics," *Pharmacol Ther*, vol. 120, no. 2, pp. 129–48, 2008.

[24] K. A. Matreyek, L. M. Starita, J. J. Stephany, B. Martin, M. A. Chiasson, V. E. Gray, M. Kircher, A. Khechaduri, J. N. Dines, R. J. Hause, S. Bhatia, W. E. Evans, M. V. Relling, W. Yang, J. Shendure, and D. M. Fowler, "Multiplex assessment of protein variant abundance by massively parallel sequencing," *Nat Genet*, vol. 50, no. 6, pp. 874–882, 2018.

[25] T. L. Mighell, S. Evans-Dutson, and B. J. O'Roak, "A saturation mutagenesis approach to understanding pten lipid phosphatase activity and genotype-phenotype relationships," *Am J Hum Genet*, vol. 102, no. 5, pp. 943–955, 2018.

[26] L. Gerasimavicius, B. J. Livesey, and J. A. Marsh, "Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure," *Nat Commun*, vol. 13, no. 1, p. 3895, 2022.

[27] S. Iqbal, E. Perez-Palma, J. B. Jespersen, P. May, D. Hoksza, H. O. Heyne, S. S. Ahmed, Z. T. Rifat, M. S. Rahman, K. Lage, A. Palotie, J. R. Cottrell, F. F. Wagner, M. J. Daly, A. J. Campbell, and D. Lal, "Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants," *Proc Natl Acad Sci U S A*, vol. 117, no. 45, pp. 28201–28211, 2020.

[28] K. Lage, "Protein-protein interactions and genetic diseases: The interactome," *Biochim Biophys Acta*, vol. 1842, no. 10, pp. 1971–1980, 2014.

[29] M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J. B. Holmes, B. L. Kattman, and D. R. Maglott, "Clinvar: improving access to variant interpretations and supporting evidence," *Nucleic Acids Res*, vol. 46, no. D1, pp. D1062–D1067, 2018.

[30] P. D. Stenson, M. Mort, E. V. Ball, K. Evans, M. Hayden, S. Heywood, M. Hussain, A. D. Phillips, and D. N. Cooper, "The human gene mutation database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies," *Hum Genet*, vol. 136, no. 6, pp. 665–677, 2017.

[31] N. Brandes, G. Goldman, C. H. Wang, C. J. Ye, and V. Ntranos, "Genome-wide prediction of disease variants with a deep protein language model," *bioRxiv*, p. 2022.08.25.505311, 2022.

[32] P. L. Martelli, P. Fariselli, C. Savojardo, G. Babbi, F. Aggazio, and R. Casadio, "Large scale analysis of protein stability in omim disease related human protein variants," *BMC Genomics*, vol. 17 Suppl 2, p. 397, 2016.

[33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv e-prints*, p. arXiv:1907.11692, July 2019.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv e-prints*, p. arXiv:1810.04805, Oct. 2018.

[35] R. P. Joosten, T. A. te Beek, E. Krieger, M. L. Hekkelman, R. W. Hooft, R. Schneider, C. Sander, and G. Vriend, "A series of pdb related databases for everyday needs," *Nucleic Acids Res*, vol. 39, no. Database issue, pp. D411–9, 2011.

[36] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–637, 1983.

[37] M. T. Chang, T. S. Bhattarai, A. M. Schram, C. M. Bielski, M. T. A. Donoghue, P. Jonsson, D. Chakravarty, S. Phillips, C. Kandoth, A. Penson, A. Gorelick, T. Shamu, S. Patel, C. Harris, J. Gao, S. O. Sumer, R. Kundra, P. Razavi, B. T. Li, D. N. Reales, N. D. Socci, G. Jayakumaran, A. Zehir, R. Benayed, M. E. Arcila, S. Chandarlapaty, M. Ladanyi, N. Schultz, J. Baselga, M. F. Berger, N. Rosen, D. B. Solit, D. M. Hyman, and B. S. Taylor, "Accelerating discovery of functional mutant alleles in cancer," *Cancer Discov*, vol. 8, no. 2, pp. 174–183, 2018.

[38] F. Cunningham, J. E. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, O. Austine-Orimoloye, A. G. Azov, I. Barnes, R. Bennett, A. Berry, J. Bhai, A. Bignell, K. Billis, S. Boddu, L. Brooks, M. Charkhchi, C. Cummins, L. Da Rin Fioretto, C. Davidson, K. Dodiya, S. Donaldson, B. El Houdaigui, T. El Naboulsi, R. Fatima, C. G. Giron, T. Genez, J. G. Martinez, C. Guijarro-Clarke, A. Gymer, M. Hardy, Z. Hollis, T. Hourlier, T. Hunt, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J. C. Marugan, S. Mohanan, A. Mushtaq, M. Naven, D. N. Ogeh, A. Parker, A. Parton, M. Perry, I. Pilizota, I. Prosovetskaia, M. P. Sakthivel, A. I. A. Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, J. G. Perez-Silva, W. Stark,

E. Steed, K. Sutinen, R. Sukumaran, D. Sumathipala, M. M. Suner, M. Szpak, A. Thormann, F. F. Tricomi, D. Urbina-Gomez, A. Veidenberg, T. A. Walsh, B. Walts, N. Willhoft, A. Winterbottom, E. Wass, M. Chakiachvili, B. Flint, A. Frankish, S. Giorgetti, L. Haggerty, S. E. Hunt, I. I. GR, J. E. Loveland, F. J. Martin, B. Moore, J. M. Mudge, M. Muffato, E. Perry, M. Ruffier, J. Tate, D. Thybert, S. J. Trevanion, S. Dyer, P. W. Harrison, K. L. Howe, A. D. Yates, D. R. Zerbino, and P. Flicek, "Ensembl 2022," *Nucleic Acids Res*, vol. 50, no. D1, pp. D988–D995, 2022.

[39] D. Esposito, J. Weile, J. Shendure, L. M. Starita, A. T. Papenfuss, F. P. Roth, D. M. Fowler, and A. F. Rubin, "Mavedb: an open-source platform to distribute and interpret data from multiplexed assays of variant effect," *Genome Biol*, vol. 20, no. 1, p. 223, 2019.

[40] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, A. d. Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives, "Language models of protein sequences at the scale of evolution enable accurate structure prediction," *bioRxiv*, 2022.

[41] J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, and D. S. Marks, "Disease variant prediction with deep generative models of evolutionary data," *Nature*, vol. 599, no. 7883, pp. 91–95, 2021.

[42] P. Notin, M. Dias, J. Frazer, J. M. Hurtado, A. N. Gomez, D. Marks, and Y. Gal, "Tranception: Protein fitness prediction with autoregressive transformers and inference-time retrieval," in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 16990–17017, PMLR, 17–23 Jul 2022.

[43] D. Stein, C. S. Bayrak, Y. Wu, P. D. Stenson, D. N. Cooper, A. Schlessinger, and Y. Itan, "Genome-wide prediction of pathogenic gain- and loss-of-function variants from ensemble learning of diverse feature set," *bioRxiv*, p. 2022.06.08.495288, 2022.

[44] M. Hu, F. Yuan, K. K. Yang, F. Ju, J. Su, H. Wang, F. Yang, and Q. Ding, "Exploring evolution-based & -free protein language models as protein function predictors," *arXiv e-prints*, p. arXiv:2206.06583, June 2022.

A Appendix

A.1 MLP layer for each task

Table 3: MLP layer for each task

Name		MLP layer
Secondary Structure		(1280, 8) Linear + Softmax
Relevant Solvent Accessibility		(1280, 1) Linear
Pathogenicity		(1280, 1) Linear + Sigmoid
Deep mutational scan	PTEN	(1280, 2) Linear
	NUDT15	(1280, 2) Linear
	CCR5	(1280, 3) Linear
	CXCR4	(1280, 3) Linear
	VKORC1	(1280, 2) Linear
GoF/LoF	Ion channel family	(1280, 2) Linear + Softmax
	SH2 family	(1280, 2) Linear + Softmax
	Kinase family	(1280, 2) Linear + Softmax
	cNMP binding family	(1280, 2) Linear + Softmax

A.2 Data set sizes for each task

Table 4: Data set sizes for each task

Name			Training set	Testing set
Secondary Structure			21,052	2,341
Relevant Solvent Accessibility			21,052	2,341
Pathogenicity	Pathogenic		58,888	877
	Benign		58,888	1,754
Deep mutational scan	PTEN	stability	3,142	392
		enzyme activity		
	NUDT15	stability	2,276	284
		enzyme activity		
	CCR5	stability	5,236	654
		bind AB-2D7		
		bind AB-12G5		
	CXCR4	stability	5,316	664
		bind CXCL12		
		bind AB-12G5		
	VKORC1	stability	582	72
		enzyme activity		
Ion channel family	GoF		249	54
	LoF		405	110
SH2 family	GoF		60	15
	LoF		20	5
Kinase family	GoF		42	11
	LoF		92	23
cNMP binding family	GoF		18	5
	LoF		58	15

A.3 Training settings

As shown by previous work that naively fine-tuning language models without regularization with the original mask-and-predict task will rapidly result in overfitting[19], we decided to train the model with freezing parameters of protein language models while update the other parameters using Adam algorithm. For pre-training, we set learning rate to 1e-5, as suggested in ESM1v paper[19], with 5 epochs of warm-up followed by 15 epochs linear learning rate decay to 5e-6. For transfer learning to protein family specific tasks, we train the model with freezing parameters of both protein language models and the transformer layer. We set the learning rate to 5e-6 with 5 epochs of warm-up followed by 15 epochs of linear learning rate decay to 2.5e-6. We set a dropout layer with rate of 0.1 after MSA-attention layer and GRU cell during training to avoid overfitting. For transfer learning tasks, we also regularize the model by keeping the pre-training loss. Furthermore, we trained the model 3 times in transfer learning and calculated the average AUC for comparison with base line methods.

A.4 Loss and training time

For secondary structure predictions, we use cross-entropy loss during training:

$$Loss = -\sum_{i,c} y_{i,c} \log(p_{i,c}),\ c \in \{-, E, H, T, S, G, B, I\} \tag{4}$$

There are 8 classes of secondary structures, including ’E’ (beta strand), ’H’ (alpha helix), ’T’ (turn), ’S’ (bend), ’G’ (3-10 helix), ’B’ (short beta bridge), ’I’ (pi helix), ’-’ (random coil). For pathogenicity predictions, we use binary-cross-entropy loss during training:

$$Loss = \sum_{i} -y_i \log(p_i) + (1 - y_i) \log(1 - p_i),\ y \in \{0, 1\} \tag{5}$$

For protein family specific gain/loss of function predictions, we use cross-entropy loss:

$$Loss = -\sum_{i,c} y_{i,c} \log(p_{i,c}),\ c \in \{gain, loss\} \tag{6}$$

Although in this case cross-entropy loss of two classes is mathematically identical to binary-cross-entropy loss, we kept using this form for easier expansion to future tasks which can have more than two modes. For relevant solvent accessibility (RSA), PTEN protein stability and enzyme activity tasks, we use mean squared error loss:

$$Loss = \sum_{i} (x_i - y_i)^2 \tag{7}$$

We used 3 NVIDIA A40 GPU for pre-train and 1 NVIDIA A40 GPU for transfer learning. The training time is estimated in Table 5.

Table 5: Training time

Task Name	Time
Pre-train	∼32h
PTEN	∼4h
NUDT15	∼2.5h
CCR5	∼6h
CXCR4	∼6h
VKORC1	∼1h
Ion Channel	∼1h
SH2, Kinase, cNMP binding families	∼0.5h

A.5 AUC and model comparison for pre-train

The AUC on testing dataset during pre-train is plotted in Figure 2. The goal of pre-train is to let the model learn representation of variant effects as well as the causal factors to benefit transfer learning while not over-fitting on the pre-train tasks. We noticed that all four models converged after epoch 8. ESM1b-based RESCVE reached the highest AUC on pathogenicity task, while ESM2 (15B)-based RESCVE reached the highest AUC on secondary structure and relevant solvant accessibility tasks. We thus selected ESM1b-based RESCVE at epoch 8 for further transfer learning tasks.

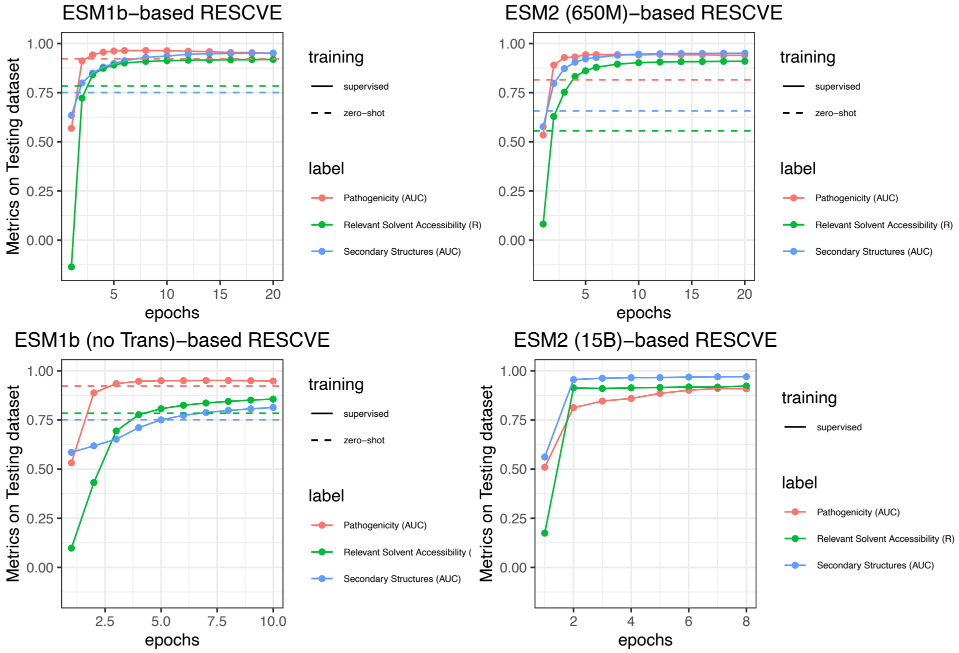

Figure 2: AUC on testing during Pre-train

A.6 Number of points for transfer-learning

We studied how RESCVE's performance would be impacted by the number of points in transfer learning. We random sub-sample 10%, 20%, 40%, 60% and 80% of each mode of action dataset as training in the transfer learning while kept same 10% of dataset as testing. We did this experiment with 3 replicates for *PTEN*, *NUDT15*, *CCR5*, *CXCR4*, *VKORC1* and Ion channel family as they have sufficient data points. We noticed that for our pre-trained RESCVE can adapt to tasks of *CXCR4* and Ion channel family with only 60% of dataset, while for *NUDT15*, 40% of data would be enough (Figure 3). For other tasks like *CCR5* and *VKORC1*, more experimental data is required for a full landscape of mutational effects on all residues.

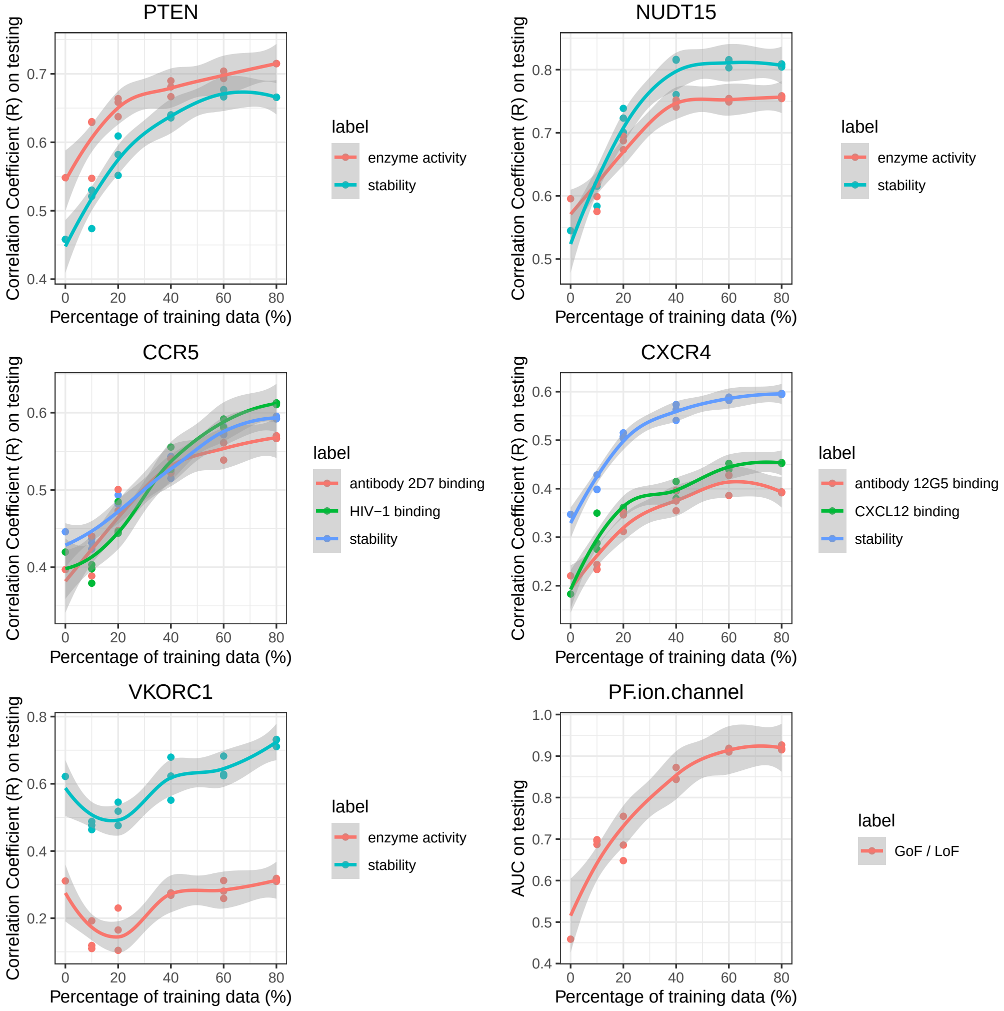

Figure 3: Number of points for transfer learning

A.7 Identify multiple functional effects of single variants

One potential application of our method is to distinguish multiple functional effects of a single missense variant. In Figure 4, we showed that RESCVE is able to distinguish disruption of protein stability and enzyme activity in the testing dataset. To be more specific, RESCVE can find mutations that decrease enzyme activity while maintain protein structure (Figure 4, bottom right corner in each plot). This will be extremely helpful for clinical diagnosis and precise treatments.

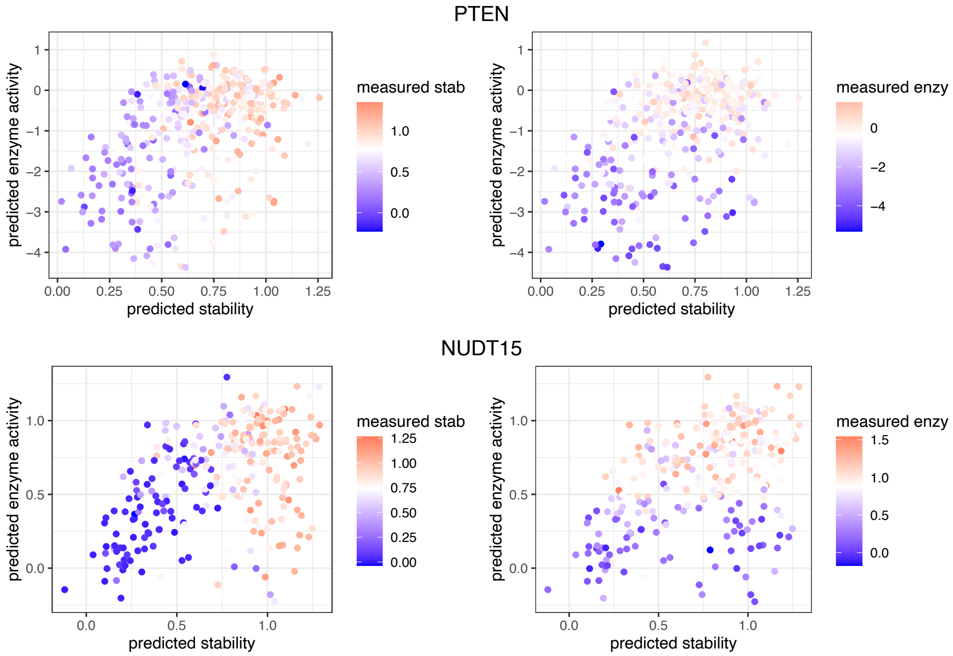

Figure 4: Multiple effects of single variants. x and y axis, model prediction; color, experimental measurements

A.8 License

The datasets used in this project are publicly available to download through the original publication except HGMD dataset. The code for this project is under GPL license and available here: https://github.com/ShenLab/rescve. We provided our pre-processed datasets (except HGMD dataset) and pre-trained models in the github repo as well.